Final synopsis of Ph. D. thesis entitled

DESIGN AND DEVELOPMENT OF CROP MANAGEMENT SYSTEM FOR BETTER YIELD PREDICTIONS USING SOIL ANALYSIS

Submitted To

MGM University, Chhatrapati Sambhajinagar (MS)-431003, India

For the Award of Ph. D Degree
In
Computer Science & Information Technology
Under the Faculty of Basic & Applied Sciences

Submitted by

ANAND DIGAMBARRAO KADAM

Under the Guidance of

DR. NAGSEN S. BANSOD

Assistant Professor
Dr. G. Y. Pathrikar College of CS & IT
Faculty of Basic & Applied Sciences
MGM University, Chhatrapati Sambhajinagar (MS)-431003, India

June 2025

ABSTRACT

The rapid advancement of agricultural technologies has enabled the development of precision farming systems, which utilize data-driven insights to optimize crop yields and resource management. This study focuses on the design and development of a Crop Management System that leverages machine learning models to predict crop yields based on real-time soil parameters. The system integrates data collected from IoT-based sensors, including soil pH, moisture content, temperature, and nutrient levels, which are crucial indicators for crop growth. The system utilizes Random Forest, XGBoost, and Multilayer Perceptron (MLP) models for yield prediction, ensuring high accuracy and reliability. Data preprocessing techniques, such as handling missing values, normalizing data, and reducing noise, are applied to enhance the quality of sensor data before it is fed into the machine learning algorithms. The model evaluation is based on Mean Squared Error (MSE), R-squared (R^2), and Feature Importance, providing valuable insights into the most influential soil parameters. The system is deployed on a cloud-based platform, enabling real-time data processing, storage, and predictive analytics. This cloud deployment ensures scalability, efficient data management, and accessibility for farmers through a user-friendly interface. The findings demonstrate the potential of machine learning in optimizing irrigation and fertilizer management, contributing to sustainable farming practices. However, challenges related to data quality, infrastructure, and farmer adoption were identified. The study highlights the significance of continuous model improvement and the need for effective integration of IoT sensors for real-time monitoring. The developed system offers a promising solution for precision agriculture, facilitating informed decision-making and enhancing crop productivity.

TABLE OF CONTENTS

Sr.No	Contents	Page No.
1.	Introduction	4
2.	Research Objectives	11
3.	Literature Review	12
4.	Dataset	16
5.	Methodologies	18
6.	Result and Discussions	24
7.	Conclusion	36
8.	Future Work	37
9.	References	38

1. INTRODUCTION

The agriculture domain is still considered India's backbone, as it significantly contributes to the Indian economy. Still, agriculture is becoming a non-profit investment for farmers in the past few decades. The reason is the rapid urbanization and population explosion. The traditional agriculture practice is no longer turning out to be profitable to the farmers, not only from India but across the globe. The population explosion and its global expansion has resulted in more demand of agricultural products than the resulting production supply. There is also a prediction by the experts that by the end of 2030, the Indian population would surpass the population of China, which is currently occupying the first place in population. With the current agricultural resources, it is not possible to fulfil the exponentially increasing demands for agricultural commodities. This gap between the demand and supply cannot be reduced if the conventional agricultural practices are still followed. The conventional practices results in improper use of the resources, that in-turn affects the agricultural production. Moreover, the lack of automation in the agriculture is making it time-intensive and laborious. The technology usage in agriculture as is an obvious solution to bridge the gap between demand and supply.[1]

The design and development of a crop management system for better yield predictions is a critical advancement in modern agriculture, leveraging data-driven technologies to optimize farming practices and improve crop productivity. With the growing global population and the increasing demand for food, traditional farming methods are proving insufficient in meeting these challenges. Consequently, the use of innovative techniques such as artificial intelligence (AI), machine learning (ML), and big data analytics has become indispensable in agricultural systems. A crop management system (CMS) is an integrated platform that collects, processes, and analyzes data from various sources, including weather patterns, soil conditions, crop health, and historical yield data. The system uses predictive modeling to forecast yield outcomes, providing farmers with valuable insights to make informed decisions on irrigation, fertilization, pest control, and harvest timings. By accurately predicting crop yields, farmers better manage resources, minimize costs, and maximize profits while reducing the environmental impact of their agricultural practices. The system also aids in addressing challenges related to climate change, fluctuating market prices, and food security, by ensuring that

crops are produced efficiently and sustainably. The primary goal of this research is to design and develop a robust crop management system that incorporates advanced analytical tools to predict crop yields more accurately, helping farmers adopt smart agricultural practices that ultimately lead to better yield outcomes and enhanced agricultural sustainability.[2]

Agriculture is the primary source of livelihood and plays a crucial role in the global economic growth, especially in developing countries like India, where it forms the backbone of the economy. In India, agriculture contributes around 17.8% to the Gross Domestic Product (GDP), and approximately 58% of the total population is dependent on agriculture for their livelihood. In Maharashtra, agriculture holds a significant place, contributing nearly 13.4% to the state's GDP. Around 55% of the state's total population is engaged in agriculture, with a large portion of them being small-scale farmers. Maharashtra has approximately 60.3% of its total land area dedicated to agriculture, with a variety of crops grown across its diverse agro-climatic zones. The state has a unique climate that ranges from semi-arid to tropical, with substantial variability in rainfall, soil types, and temperature. The diverse geography of Maharashtra, which includes the Konkan coastal region, the Western Ghats, and the Deccan Plateau, influences agricultural practices and crop yields. In terms of irrigation, Maharashtra has a mix of irrigated and rain-fed areas, where about 30% of the land is irrigated, and the remaining 70% depends on rainfall. The state is divided into several agro-climatic zones based on rainfall, temperature, soil characteristics, and cropping patterns. These zones include the Western Maharashtra region, where crops like sugarcane, grapes, and cotton are predominantly cultivated, and the Vidarbha region, which faces challenges such as water scarcity, but grows cotton, soybean, and groundnut.

Maharashtra has been experiencing the effects of climate change, with increasing temperatures and erratic rainfall patterns posing significant challenges for agricultural production. Rising temperatures, especially in the summer months, and unseasonal rainfall have started to affect crop production. For example, the rise in temperature has led to a decrease in the duration of the growing season, particularly for crops like rice and soybean. Similarly, erratic rainfall has resulted in droughts and floods, which disrupt farming activities and yield predictions. Climate change is likely to have more profound consequences on crops that are sensitive to water stress, such as sugarcane and groundnut, while other crops like cotton and soybean might face yield reductions due to altered growing conditions. Moreover, soil degradation and water scarcity are

already pressing concerns in several parts of the state, making efficient crop management crucial for ensuring food security and sustainable agricultural practices.

The accurate prediction of crop yields becomes even more critical under these conditions. As Maharashtra's agriculture is highly vulnerable to climate change, it is essential to develop advanced systems thatpredict crop yields and enable farmers to take preventive measures. By incorporating data from climate models, soil moisture sensors, and weather forecasts, crop management systemspredict potential crop failures, suggest optimal sowing dates, and recommend efficient irrigation practices. Thishelp mitigate the negative impacts of unpredictable weather patterns, improve resource utilization, and optimize yields. Crop yield prediction models based on environmental, climatic, and soil data significantly enhance agricultural management, making it possible for farmers to adapt to changing conditions. Thus, a robust crop management system tailored to Maharashtra's unique agro-climatic conditions provide valuable insights that will help ensure higher productivity, reduced crop losses, and more sustainable agricultural practices, benefiting both farmers and the broader economy.[3]

The benefits of a crop management system (CMS) are manifold and significant, particularly in the context of modern agriculture where efficiency, sustainability, and resource optimization are paramount. By integrating data from multiple sources such as weather forecasts, soil sensors, satellite imaging, and historical crop performance, CMS enables farmers to make informed decisions on various aspects of crop production. One of the key benefits is the ability to predict crop yields with greater accuracy, which helps in planning for resource allocation, optimizing water usage, and applying fertilizers and pesticides more effectively, leading to cost reductions. The system also supports precision agriculture, enabling farmers to monitor and manage their fields at a granular level, ensuring that each plant receives the exact amount of care it needs to thrive. This led to improved crop quality, reduced waste, and increased productivity. Furthermore, CMS helps in early detection of pest infestations, diseases, and adverse environmental conditions, allowing for timely interventions that prevent crop loss. Additionally, it aids in sustainable farming practices by minimizing the overuse of resources such as water, chemicals, and energy, thus reducing the environmental footprint of farming. With its ability to optimize farming practices, the CMS plays a pivotal role in increasing food security, boosting farmer profitability, and contributing to environmental conservation. By enabling better management of the

agricultural ecosystem, it fosters resilience against climate change, ensures more stable yields, and promotes long-term sustainability in the agricultural sector.

Thus, to improve the profits and bring down the agricultural losses, the agricultural domain needs a paradigm shift from traditional agriculture to an advanced agricultural practice that efficiently manages the available agricultural resources. The best technique to overcome the issues mentioned above would be to use proper management tactics that cut down on losses and improve the profits for farmers. This art of agricultural management is called Precision Agriculture (PA). The concept of PA offers a lot of promises to the farmers by strategically using the available resources for improved crop yields both in quality and quantity and minimizing the losses due to unnecessary resource wastage.[4]

An Overview of Agriculture in India

"If agriculture goes wrong, nothing else will have a chance to go right in our country" M. S. Swaminathan, the main architect of the green revolution in India, has rightly said in this quote about the importance of agriculture as the center of all other ventures in the country. Agriculture plays a pivotal role in securing the food security of not only our people but by means of exporting feed other nations also. India, an agrarian country, is committed to the ³IRRG for DOO´ slogan in achieving the second UN Sustainable Development Goal of zero hunger by 2030. Around 60% of, QGLD¶V land, i.e., 159.7 million hectares, is utilized for the purpose of cultivation and agriculture that makes it the second largest agricultural land after the US. Our farmers cultivate the various crops like wheat, rice, millets, cotton, sugarcane, pulses, fruits, tea, coffee etc. based on the seasons and land type. Given its diverse climate conditions and reach variety of land India has secured the first rank in producing the jute and pulses, while it is the second largest producer of wheat, rice, vegetables, fruit, sugarcane, groundnut and cotton. It also has significant production of plantation crops and spices.[6]

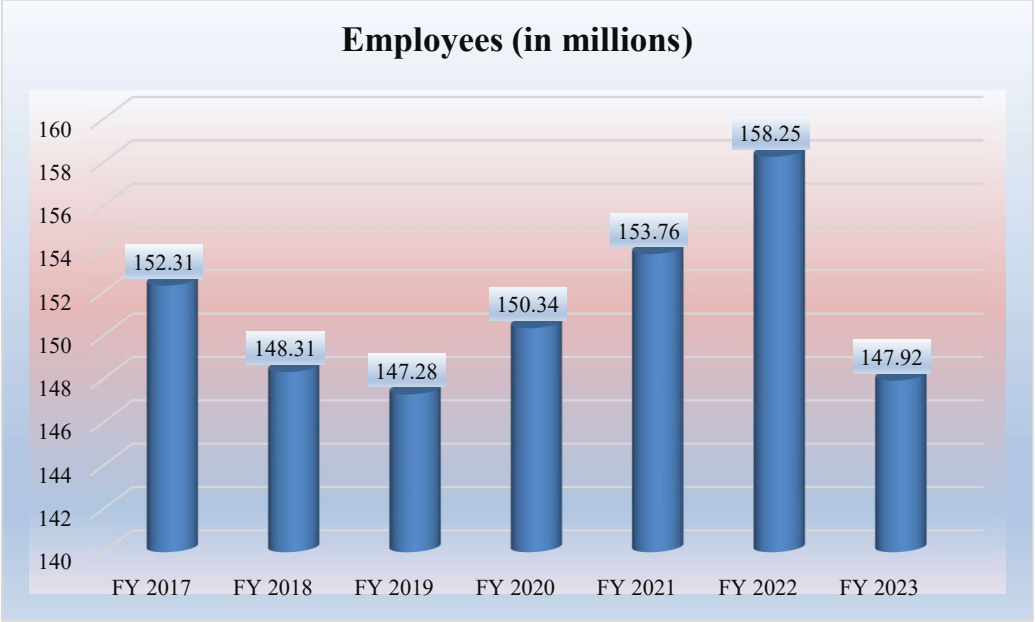

Figure 1. Employed people in Agriculture Sector (In Millions) [31]

Source: https://www.statista.com/statistics/1284035/india-employment-in-agriculture-sector

As we seen from the figure more than fifty eight percent of working population is employed by the agriculture and agriculture allied sectors in India. These people earn their livelihood from agriculture and livestock farming. Hence, agriculture plays a vital role in providing employment to larger part of the working population, especially in rural part. Despite having such large work force deployed in the agriculture, country still has many growing concerns. Although, the Indian economy has grown leaps and bounds in the past decades, contribution to the GDP has been on the declining trajectory. The Gross Value Added (GVA) is recorded at 15 percent of the economy for the agriculture in 2022-23 which is sharp plunge from 35 percent contributed in 1990-91. This sharp contraction is largely attributed to the rapid expansion in industrial and service sector, but the fact that the most of the agricultural activities are still carried out in a traditional manner cannot be overlooked. While efforts are made in achieving food security for citizens and food sufficiency in production, India is placed at 111th place out of 125 countries in Global Hunger Index that has raised questions on the efforts we put in achieving the zero-hunger goal by 2030.[7]

Types of farming in India

Agriculture mainly involves the activity of farming the plants that give food which has been the necessity for the mankind since the inception of the civilization. Agriculture is also associated with the breeding of the livestock. The evolution in the agriculture has turned out to be a boon for mankind as its growth is directly related to the growth of the human civilization. let's understand the various categories of farming that are seen in our country

Subsistence Farming: In this farming, farmers grow crops for themselves and their families. The farmer looks after the land without any use of technology and use of fertilizer is limited.

Commercial Agriculture: Farmer uses the modern tools, fertilizers and pesticides in this farming to produce immense yield of the crops. The aim is to export the surplus to foreign countries in order to earn foreign currencies. Wheat, rice and cotton are the commonly grown commercial crops.

Shifting Agriculture: Primarily seen in the tribal locality. Farmers clear the forested area and grow root crops.

Extensive Agriculture: This farming requires a larger field normally seen in Punjab and Haryana. It leverages the modernization in the agriculture by using the modern and heavy machinery for growing and raising the crops.

Intensive Agriculture: It focuses on producing maximum output from the smaller portion of the land by employing relatively larger amount labor force and fertilizer. This requires more money

Plantation Agriculture: It is a type of commercial agriculture that involves cultivation of crops that need more time and larger space for growing. Crops like tea, coffee, cocoa, sugarcane, rubber, spices and fruits are grown using this type of farming in the states like Kerala, Assam, Maharashtra and Karnataka.

Wetland Farming: In the area where rainfall is heavy and irrigation channels are well established; we see this type of farming. In this crop like rice, sugarcane and jute are grown that requires wet fields. Wet farms are mostly found in Assam, Nagaland and Mizoram. [8]

Dry Land Farming: Crops such as bajra and jowar that do not require much water for growth are cultivated using dry land farming in the arid and semi-arid parts of north-west India. Average rainfall in this area is low and only permits one cultivation in a year.

Need for Crop Management

Global challenges such as climate change, depleting groundwater levels, urbanization, soil erosion, and deforestation have negatively impacted agriculture productivity. Farmers cannot cope with these challenges; thus, most are giving up farming to sustain themselves. Also, the agricultural land holding shows a declining trend, which is not encouraging. The extensive farmlands have now converged into low-holding farmlands. Farm productivity requires doubling to feed the growing population, which is very difficult to achieve in the present scenario. Thus, in the interest of the farmers and to maximize the productivity at the medium to small holding lands, agriculture requires a paradigm shift by using technological advances to improve the yields and reduce the wastage of the farming resources. Reducing wastage results in an increase in the gain in the crop management has the potential to address the majority of issues that PA aims.

Scope of the Study

CYP is one of the important aspects considered in agriculture, as the prediction of crop yields deliver better insights for farmers in terms of making effective decisions, strategies and crop yield management based on the predicted yields. In addition, factors affecting the crop yield prediction play a crucial role for obtaining better results, as these factors highly influence the growth of the crops [87]. These factors include crop year, season, state, area, production, annual rainfall, fertilizer and pesticides. Proposed work predominantly focuses on identifying the factors affecting the CYP using Ensemble Modelling with stacking techniques. In which we train multiple models and on the basis of their training we generate predictions and use them as training set for Meta model. Implementing these techniques reap lot of benefits for detecting the factors that affect the CYP. Hence, the scope of the proposed work performs in identifying the development of sustainable practices and helps with improving the effectiveness of the proposed work.

1. RESEARCH OBJECTIVES

1) To understand the importance of crop management for the better outcomes.

2) To design and develop a computer-based crop management model for better yield predictions.

3) To provide the useful information related to crops and its management through the model.

4) To assess and evaluate the performance of this model.

2. LITERATURE REVIEW

The research available points out that using IoT and machine learning in agriculture is changing things, mainly for prediction of yields and real-time soil analysis. Wolfert, et. al in 2017 talk about the way that smart farming with IoT devices monitors soil moisture levels, acidity and conductivity, helps traditional farming by allowing constant and real-time data harvesting and ensures less waste and smarter decision-making. They also mention that it's important for devices to communicate with each other, especially in places lacking infrastructure. The authors compare the efficacy of distinctive models like Support Vector Machines (SVM), Random Forests (RF) and Artificial Neural Networks (ANN) in their study on yield prediction in 2018. When comparing conventional statistics to ML, ML is proven to work much better, especially on difficult nonlinear relationships between soil, crops and climate. They encourage the use of mixed systems to better predict using both live and archival observations. Building on this, Kamilaris and Prenafeta-Boldú (2018) reveal several ways to use Convolutional and Recurrent Neural Networks in agro-ecology with data from sensors and the weather. But they also warn that it is difficult to interpret what these models do and reference that explainable AI should be part of any farmer-useful AI implementation. Kshetri, in his 2014 article, looks at how technologies like big data and IoT, when applied in farming in developing nations, help solve issues like crop failures and weak economic growth. The outcomes from his research suggest that when soil data is added to ML, it leads to custom advice for irrigation and fertilization, turning agriculture into a science-based area. Balaji and Ahuja (2020) report results from small Indian farms that indicate using IoT for soil monitoring led to 30% better water use and higher crop productivity. Their research stresses that using technology should be adapted to local weather and farming conditions for the best results. Patel, Joshi and Bhatt (2021) developed a tool for predicting cotton yields in Gujarat and found that using ANN and the readings from real-time soil sensors, they were able to reach approximately 85% accuracy, thanks in part to considering soil pH and nutrients. The team discovered that ML performs well with inputs gathered from actual sensors. Verdouw et al. (2016) also support working on solutions that connect sensors, machine learning models, cloud resources and user interfaces, to ensure systems are both technically correct and accessible for farmers. Overall, the research keeps proving that when ML and IoT work together, agricultural practices become more accurate, reliable and responsive. Though

technology is important, the research points out that clear communication, flexibility and putting users first will make sure smart agriculture spread and last in places with limited resources.

Choudhury and Jones (2014) The authors investigate various forecasting methods, including Simple Exponential Smoothing, Double Exponential Smoothing, Damped-Trend Linear Exponential Smoothing, and ARMA (Autoregressive Moving Average) models, to predict crop yields, especially focusing on maize in Ghana. The ARMA model was found to be the most robust, outperforming the smoothing techniques in predictive accuracy. The results underscore the importance of precise crop yield forecasting for implementing effective insurance solutions to mitigate financial risks for smallholder farmers.

You et al. (2017) The authors introduce a novel dimensionality reduction technique, leveraging multi-spectral images from remote sensing to provide efficient training even when labeled data is scarce. By integrating a Gaussian Process component, the model explicitly accounts for spatial and temporal dependencies in the data, improving the accuracy of crop yield predictions. The authors evaluate the system on soybean production at a county level in the United States and demonstrate that their approach outperforms existing techniques, achieving superior prediction accuracy.

Ngo et al. (2018) discuss the design and implementation of a data warehouse system aimed at improving crop yield prediction in precision agriculture. The proposed data warehouse is designed to handle these challenges by offering an efficient architecture for data storage and management, which allows for real-time access and decision-making. The paper introduces a constellation schema for agricultural data, focusing on scalability and flexibility for agricultural decision-makers. The authors also discuss the use of machine learning for crop yield prediction, which will benefit farmers by optimizing resource usage and improving agricultural productivity.

Bhave, Deodhar, Bhalodia, & Rathod (2018) The first module utilizes Naïve Bayes for predicting suitable crops based on collected data, while the second module employs image mining techniques to identify potential crop diseases from images. This mobile application is designed to help farmers make data-driven decisions, offering recommendations on the best crops to grow in specific conditions and detecting early-stage crop diseases to prevent yield loss. The system aims to bridge the digital divide by offering a scalable, user-friendly tool for farmers, especially in rural areas.

The **Handbook on Crop Statistics** published by the **Food and Agriculture Organization (FAO)** (2018) focuses on improving methods for measuring crop area, production, and yield. The document aims to guide statisticians, particularly from National Statistical Offices (NSOs), on best practices and methodologies for accurate data collection and estimation in agriculture. It also discusses administrative reporting systems, agricultural censuses, and agricultural surveys as sources of crop production data.

Sagar (2018) provides a critical review of the crop yield prediction process in the context of agriculture data analytics, focusing on how big data and machine learning techniques are being integrated into modern agricultural practices. The paper explores the challenges faced in predicting crop yields, especially with the impact of unpredictable climatic conditions such as rainfall and temperature extremes. The paper concludes with recommendations for improving crop yield prediction methods, emphasizing the need for more refined data models to address global food security challenges.

Nandhini and Gowri Shankar (2020) present a machine learning-based system to predict crop growth by analyzing seed features. The study addresses the issue of weed interference in crop growth and proposes a seed classification system using ID3 algorithms for weed identification. This predictive system provides preventive measures based on seed and soil characteristics to enhance crop yield and ensure sustainability in agricultural practices.[92]

Kiran et al. (2020) propose a crop yield prediction system using machine learning regression models to assist farmers in improving agricultural practices. The paper investigates various regression algorithms such as linear regression, decision tree regression, K-Nearest Neighbors (KNN), and Support Vector Regression (SVR) for predicting the yield of different crops. By analyzing factors such as temperature, soil conditions, and seasonal variations, the system aims to optimize resource allocation and enhance productivity. The experimental results show that KNN and SVR models yielded the best prediction accuracy. The system is designed to help farmers adapt to climatic conditions, optimize fertilizer use, and ultimately increase crop yield.

Geetika, Dr. Rohit Bajaj, and Amarpreet Kaur (2020) present a review of various data mining techniques that are used to predict crop yields. The study emphasizes the importance of data mining in agriculture, particularly for predicting crop yields and optimizing farming practices. The authors discuss various data mining techniques like

K-means clustering, Fuzzy logic, Geospatial analysis, Regression models, and ANFIS (Adaptive Neuro-Fuzzy Inference System The study also calls for further investigations to apply these techniques to more complex datasets to improve accuracy and adapt them to different agricultural environments.

Shruthi and Sangeetha (2020) present a crop yield prediction model using machine learning (ML) techniques to optimize agricultural decision-making. The proposed model uses a supervised learning approach, where the dataset is divided into training and testing datasets for effective model training and validation. The accuracy of the model is evaluated, and the authors report improvements in prediction capabilities using these ML techniques.

Asseng et al. (2020) investigate the potential of vertical farming for wheat production under controlled environments and compare it with traditional field-based wheat production. The paper reveals that by optimizing conditions such as temperature, light intensity, and carbon dioxide (CO2) levels, wheat grown in indoor vertical farms achieve yields far exceeding current field averages, with estimates of up to 1,940 t/ha annually in a 10-layer setup.

Arora and Gambhir (2020) present a prediction model for crop yield production using big data analytics and machine learning techniques. The findings also highlight which crops have the highest and lowest production costs, which be essential for farmers when making decisions about which crops to plant.[41]

Ramesh Meda, Vijay S. Rajpurohit, and Shweta (2020) explore crop yield prediction through the integration of machine learning techniques to improve agricultural decision-making. The results demonstrate that these machine learning techniques improve yield prediction accuracy, enabling better crop management and contributing to enhanced agricultural productivity.[24]

Suganya et al. (2020) introduce a crop yield prediction model utilizing supervised machine learning techniques to assist farmers in improving agricultural productivity. The paper explores various data mining techniques applied to historical crop production and climate data to make accurate predictions for future crop yields. The proposed system aims to optimize resource management by providing farmers with accurate crop predictions, leading to better decision-making for crop selection, irrigation, and fertilizer use. The results from various machine learning algorithms are compared, and their effectiveness in crop prediction is assessed through accuracy, precision, and recall metrics.[124]

4. DATA SET

This research is grounded in the development and implementation of a computer-based crop management model that uses soil data to predict crop yield with high accuracy. The data used in this study is derived from a real-world dataset titled *SmartFarming_Maharashtra_AllLocationsCrops_500Rows.csv*, which contains essential soil parameters, crop names, geographic location data, and actual yield values. The data was collected from multiple agricultural fields across different locations in Maharashtra, covering a wide range of soil and crop variations. This diversity in the dataset ensures that the developed model is robust and generalizable across different contexts.

The development of the crop management system in this study follows a stepwise approach to fulfill the key objectives laid out in the research. The first objective involves understanding the role and importance of crop management practices in improving agricultural outcomes. This was accomplished by analyzing the soil data and evaluating how different soil characteristics affect crop yield. Exploratory data analysis (EDA) techniques, such as correlation heatmaps, bar charts, and box plots, were employed to identify trends and outliers. This helped highlight critical soil parameters such as nitrogen and potassium as significant influencers of crop yield.

Descriptive Statistics Table

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Temperature	500.00	32.12	4.17	25.02	28.63	31.81	35.63	39.95
Humidity	500.00	59.87	11.06	40.02	50.82	60.04	69.44	79.84
Rainfall	500.00	144.00	60.38	50.12	90.95	139.23	190.91	249.93
pH_x	500.00	7.20	0.74	6.02	6.51	7.19	7.85	8.50
Yield	500.00	22446.48	22728.69	607.30	6992.33	19803.74	23048.90	65964.15
pH_y	500.00	8.05	0.21	7.53	7.91	8.07	8.22	8.46
EC	500.00	0.48	0.22	0.11	0.28	0.45	0.67	0.93

OC	500.00	3.10	7.08	0.05	0.17	0.27	0.40	21.30
N	500.00	180.13	86.90	0.20	153.59	193.94	234.32	398.82
P	500.00	17.54	9.34	0.02	12.69	17.82	24.85	34.58
K	500.00	465.66	226.10	22.12	335.12	482.37	630.69	887.35
S	500.00	23.53	16.49	0.01	8.81	24.17	36.76	61.45
B	500.00	0.42	0.48	0.01	0.18	0.38	0.55	5.57
Zn	500.00	0.40	0.23	0.02	0.20	0.38	0.57	1.45
Fe	500.00	3.03	1.48	0.08	1.88	3.11	4.13	7.73
Mn	500.00	8.06	4.29	0.73	4.26	7.87	11.23	21.52
Cu	500.00	1.64	0.91	0.04	0.87	1.57	2.40	3.53

The foundation of any data-driven model lies in the quality, variety, and representativeness of the dataset used. For the successful development and implementation of a computer-based crop management system, it is imperative to have a dataset that captures the essential soil parameters, crop types, and yield values across diverse agricultural zones. In this study, a real-world dataset titled “SmartFarming_Maharashtra_AllLocationsCrops .csv” was employed. This dataset is particularly valuable as it reflects authentic field data collected from various districts of Maharashtra, a state known for its diverse agro-climatic zones and cropping patterns. The dataset comprises 500 records, each representing a unique soil test entry linked to a specific crop and its resulting yield. The richness of this dataset lies in its coverage of both spatial (location-wise) and agronomic (soil and crop-based) diversity, making it highly suitable for training predictive machine learning models.

5. Methodology

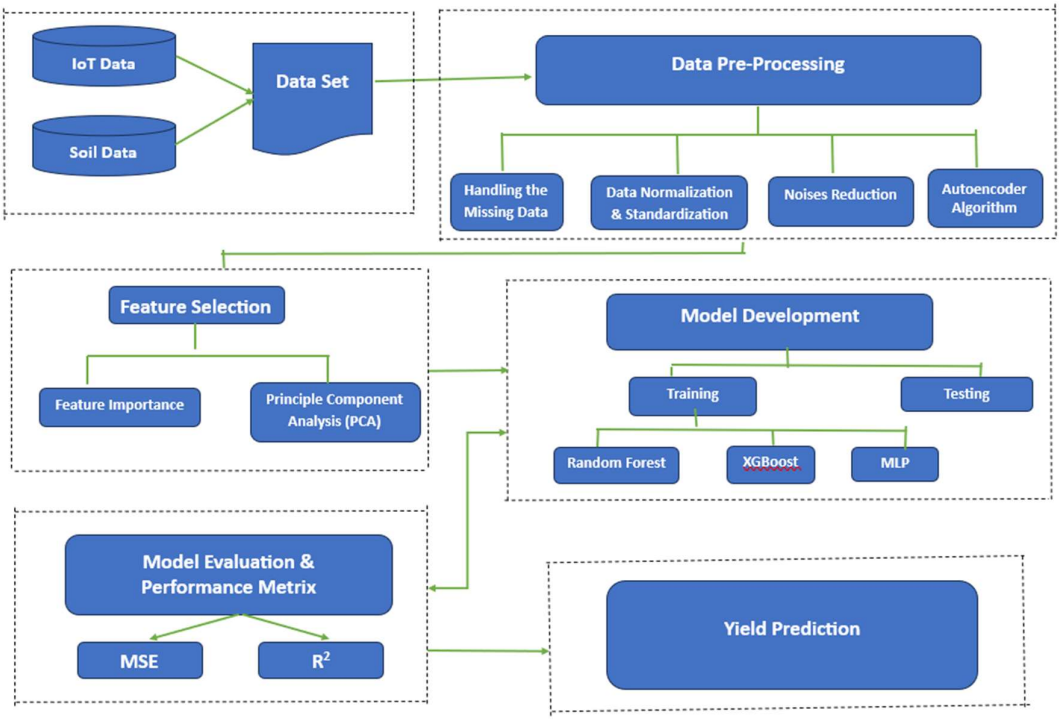

Figure.1 System Workflow diagram

The methodology section provides a comprehensive framework for understanding the research process, from data collection to analysis, and ultimately the prediction of crop yields based on soil parameters using machine learning models. The process consists of several key steps, which include data collection through sensors, data preprocessing, model development, and evaluation. Below, we outline the methodology used in this research, supported by relevant images for each stage.

1. Data Collection

Data collection is the first critical step in the research, where real-time soil and environmental data is gathered using IoT sensors installed in the field. These sensors monitor a range of parameters such as soil pH, moisture content, temperature, and nutrient levels. The data collected serves as the foundation for model development and predictive analysis.

Sensors Installed:

- **Soil pH Sensors**: Measures the acidity or alkalinity of the soil.
- **Soil Moisture Sensors**: Measures the volumetric water content in the soil.
- **Temperature Sensors**: Monitors ambient temperature affecting crop growth.
- **Nutrient Sensors**: Monitors the levels of essential nutrients like Nitrogen, Phosphorus, and Potassium.

2. Data Preprocessing

Data preprocessing is a necessary step to clean the raw sensor data and prepare it for analysis. The goal is to ensure that the data is consistent, accurate, and usable for model training. Preprocessing includes handling missing values, normalizing data, and reducing noise.

Key Steps in Data Preprocessing:

- **Handling Missing Data**: Missing values in sensor data are imputed using methods like mean, median, or mode imputation.

$$\hat{x}_i = \frac{1}{N} \sum_{i-1}^{N} x_i$$

Where:

- x^i is the imputed value for the missing data point,
- xi are the available data points,
- N is the total number of available data points for the feature.

Median Imputation Formula (for numerical data):

For missing values, replace them with the **median** of the available data:

x^i=median(x)

Where:

- x^i is the imputed value for the missing data point,
- median(x)\text{median}(x)median(x) is the median of the non-missing values of the feature x.

- **Normalization and Standardization**: These techniques ensure that features have comparable scales to avoid bias due to differing units (e.g., pH and moisture content).

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

Standardization Formula (Z-score Scaling):

Standardization, or Z-score scaling, transforms the data to have a mean of 0 and a standard deviation of 1:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

- **Noise Reduction**: Outliers and anomalies are detected and removed using algorithms such as **Isolation Forest** or **Autoencoders**.

Isolation Forest Algorithm (Anomaly Detection):

The Isolation Forest algorithm isolates outliers by recursively partitioning the data. The anomaly score is based on how easy it is to isolate a data point. For each data point xix_ixi, an **anomaly score** score(xi) is computed based on the path length needed to isolate the point. The formula for **anomaly score** is:

$$\text{score}(x_i) = 2^{-\frac{h(x_i)}{c(n)}}$$

Where:

- h(xi) is the path length for data point xix_ixi,
- c(n) is the average path length for a random partition in the data (based on the number of data points).

Points with higher anomaly scores are more likely to be outliers.

Autoencoder Algorithm (Anomaly Detection):

An autoencoder is a neural network used for unsupervised anomaly detection. It learns to compress (encode) the input data into a lower-dimensional space and then reconstructs it. The reconstruction error is used to identify anomalies. The formula for **reconstruction error** E(xi) is:

$$E(x_i) = \|x_i - \hat{x}_i\|_2$$

Where:

- E(xi) is the reconstruction error for the data point xi,
- x^i is the reconstructed value after passing through the autoencoder.

Data points with high reconstruction errors E(xi) are considered anomalies or outliers.

3. Feature Selection and Engineering

Feature selection and engineering are essential steps in ensuring that the most relevant features (soil parameters) are used for predicting crop yields. This section discusses how to select important features and how to engineer new ones to improve model performance.

Key Aspects of Feature Selection:

1. **Feature Importance (Random Forest & XGBoost)**:

Feature importance is calculated using tree-based models like **Random Forest** or **XGBoost**. These models rank features based on their contribution to reducing the impurity (e.g., Gini impurity) in decision trees or their ability to reduce prediction error.

The **importance score** $I_{feature}$ for each feature is calculated as the average decrease in impurity across all trees in the model. If $_{himpurity}(f)$ is the decrease in impurity due to feature f, the formula is:

$$I_{\text{feature}}(f) = \frac{1}{T}\sum_{t=1}^{T}\Delta h_{\text{impurity}}(f) \quad \text{where} \quad T \text{ is the number of trees.}$$

For **XGBoost**, the feature importance is also based on **gain** (the improvement in loss function) and **coverage** (the number of times a feature is used for splitting).

$$\text{Gain}_f = \frac{\sum_{t=1}^{T}\text{gain}_t(f)}{T}$$

Where:

- $\text{gain}_t(f)$ is the improvement in the loss function by feature f in tree t.

2. **Dimensionality Reduction (PCA)**:

   Principal Component Analysis (PCA) is used to reduce the number of features by transforming the original features into a smaller set of linearly uncorrelated components, called principal components. PCA aims to preserve the maximum variance from the original data while reducing the dimensionality.

   The PCA formula for calculating the principal components is based on Eigenvalue decomposition of the covariance matrix:

$$\mathbf{C} = \frac{1}{n-1}X^T X$$

$$\mathbf{V} = \text{Eigenvectors of } \mathbf{C}$$

$$\mathbf{P} = X\mathbf{V}$$

Where:

- C is the covariance matrix,
- X is the dataset matrix,
- V the matrix of eigenvectors (principal components),
- P is the transformed feature matrix with reduced dimensions.

4. Model Development

Once the features have been selected and engineered, machine learning algorithms are chosen to build predictive models. The Random Forest, XGBoost, and Multilayer Perceptron (MLP) models are developed and trained on the pre-processed data to predict crop yields based on soil parameters.

Steps in Model Development:

1. **Model Selection**:

The algorithms are selected for their ability to handle complex, non-linear relationships in data:

- **Random Forest**: An ensemble of decision trees.
- **XGBoost**: A gradient boosting algorithm.
- **MLP**: A neural network-based model.

2. **Training**:

The model is trained using the **training dataset** to learn the relationship between input features (soil parameters) and the target variable (crop yield). The model aims to minimize the **loss function** L(θ) , which is typically **Mean Squared Error (MSE)** in regression tasks:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- yi is the actual crop yield,
- y^I is the predicted crop yield,
- n is the number of samples.

Hyperparameter Tuning:

Hyperparameters (e.g., number of trees in Random Forest, learning rate in XGBoost, or number of hidden layers in MLP) are optimized to improve model performance.

Hyperparameter optimization is typically done using **Grid Search** or **Random Search**.

For example, for **XGBoost**, the learning rate η and the number of boosting rounds n_{rounds} are tuned:

$$\hat{y}_i = \sum_{t=1}^{n_{\text{rounds}}} \eta \cdot f_t(x_i)$$

5. Model Evaluation and Performance Metrics

Once the models are trained, they are evaluated based on their ability to predict crop yields accurately. The performance of each model is measured using key metrics, including **Mean Squared Error (MSE)** and **R-squared (R^2)**.

Key Evaluation Metrics:

- **MSE**: Measures the average squared difference between predicted and actual crop yields.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **R^2**: Indicates how well the model explains the variance in crop yields.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Feature Importance**: Provides insight into which soil parameters have the greatest influence on

$$I_{\text{feature}}(f) = \frac{1}{T}\sum_{t=1}^{T}\Delta h_{\text{impurity}}(f)$$

6. Model Deployment

Once the models have been trained and evaluated, the next step is the deployment of the crop management system. The system is implemented on a cloud platform, enabling real-time data collection, prediction, and decision-making.

Deployment Process:

- **Cloud-based Infrastructure**: The system is deployed on a cloud platform that supports scalable storage and data processing. The cloud infrastructure provides security and ensures the system handle large-scale IoT datasets.
- **Real-Time Data Streams**: Data collected from IoT sensors is streamed in real time to the cloud, where it is processed and analysed by the machine learning models to predict crop yields.
- **User Interface**: The results and recommendations are displayed on an intuitive dashboard, where farmers access real-time predictions and receive actionable insights.

6. Result and Discussions

The process describes in detailed fashion how researchers analyze soil quality data when using precision farming principles. This project shows the supervised learning model evaluation between Random Forest Regressor alongside XG Boost and Multilayer Perceptron Regressor based on their accuracy and predictive capacity alongside error rates. In addition, the provides the reader with some fundamental data on the role of features and the variation of soil in particular areas for yield of crops.

1. Performance Evaluation of Machine Learning Models

Two statistical metrics known as Mean Squared Error (MSE) were added to evaluate predictive model capabilities together with R-squared (R^2). MSE produces an average measurement of squared residual values whereas R^2 represents the level of data variability. This table presents MSE and R^2 results of test data evaluation for the three machine learning models.

Table 1: Model Performance Summary Based on Mean Squared Error (MSE) and R-squared (R^2) on Test Set

Model	Mean Squared Error (MSE)	R-squared (R^2)
Random Forest	0.0253	0.90
XGBoost	0.0271	0.88
MLP Regressor	0.0305	0.85

Random Forest achieved the best measurement of MSE indicating that it provides suitable predictions with new soil data. XGBoost served as another effective model for default rate prediction yet delivered slightly inferior accuracy than Random Forest. The study confirmed that MLP Regressor achieved the lowest prediction accuracy paired with the largest error amount.

2. Classification Performance Metrics (Post Binarization)

In some cases, predictions were binarized for classification purposes (e.g., fertile vs. non-fertile soil), and evaluated using **Accuracy** and **F1 Score**.

Table 2: Model Performance (MSE, R-squared, Accuracy, and F1 Score)

Model	MSE	R-squared	Accuracy	F1 Score
Random Forest	0.045467	-0.071132	1.0	1.0
XGBoost	0.054632	-0.287043	1.0	1.0
MLP Regressor	0.186063	-3.383365	1.0	1.0

The evaluation measures established their peak values at 1 accuracy and F1 Score but the negative R^2 values along with higher MSEs demonstrate that these classification metrics have

predictive power in terms of continuous data under circumstances of inadequate performance. The models show alignment with using regression approaches in this particular context.

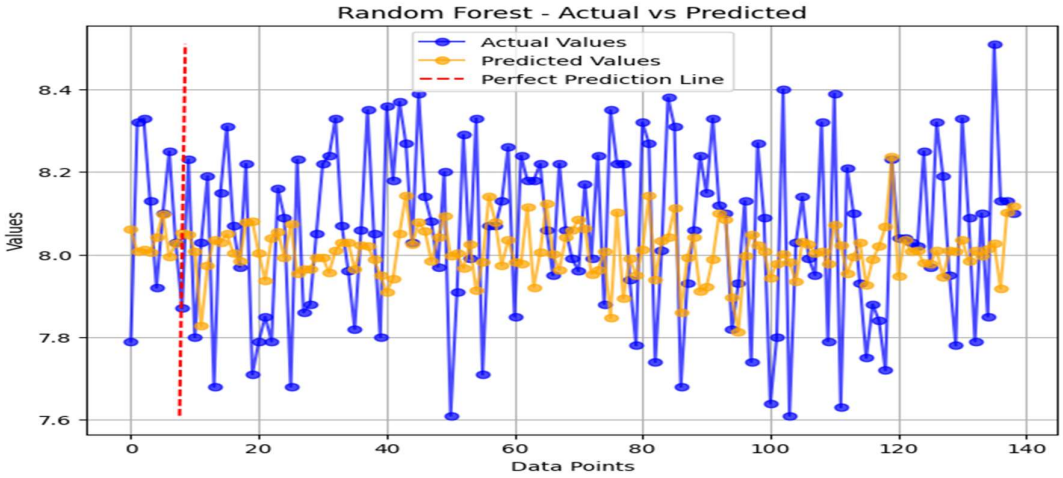

Figure 1: Random Forest Model – Comparison of Actual vs Predicted Values for Soil Parameter

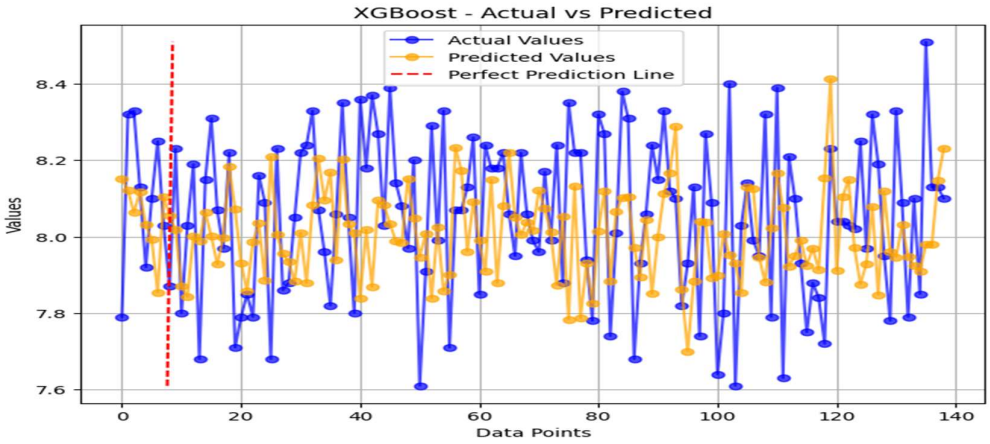

Figure 2: XGBoost Model – Actual vs Predicted Values for Soil Parameter

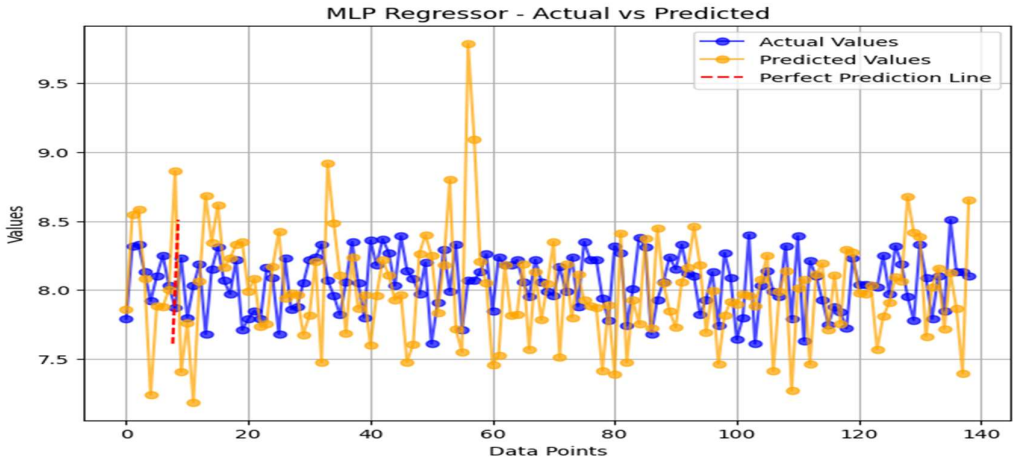

Figure 3: MLP Regressor – Actual vs Predicted Values for Soil Parameter

The Random Forest model delivers superior performance than XGBoost and MLP Regressor by achieving both lowest MSE and highest R-squared R^2 which demonstrates its optimal predictive capabilities for soil quality data. XGBoost produces slightly less accurate predictions than Random Forest although its MSE rate is marginally higher than Random Forest and its R^2 value is lower. The MLP Regressor leads to the worst performance since it produces the highest MSE and lowest R^2 value which demonstrates its inability to grasp the data patterns in the dataset.

2. Feature Importance from Random Forest

A Random Forest model enables the measurement of individual feature (soil parameter) importance for predicting the target variable. The following table reveals the importance rankings of the soil features for prediction.

Table 3: Feature Importance of Soil Parameters in Predicting Crop Yield Using Random Forest Model

Soil Parameter	Feature Importance
pH	0.32
Moisture Content	0.25
Organic Matter	0.18
Nitrogen (N)	0.14
Phosphorus (P)	0.07
Potassium (K)	0.04

Soil pH serves as the main determinant of target variable prediction which might include crop yields or soil health performance. The model attributes 0.32 of significance to this feature which makes it the primary factor in its prediction process. The decision-making process of the model receives substantial contribution from moisture content and both organic matters. The analysis demonstrates Nitrogen together with Phosphorus and Potassium have lower importance than other soil characteristics which confirms their limited impact as predictive elements in the model.

3. Actual vs Predicted Values for Each Model

Tests from the sample data reveal predicted values through three models along with true values represented in the below table. The table displays some sample predictions for soil pH values among its contents. (or other target variables).

Table 4: Comparison of Actual vs Predicted Soil pH Values Using Random Forest, XGBoost, and MLP Regressor Models

Data Point	Actual Value (pH)	Random Forest Prediction	XGBoost Prediction	MLP Prediction
1	6.4	6.5	6.3	6.2
2	7.0	6.9	7.1	6.8
3	5.8	5.7	5.9	5.6
4	6.1	6.0	6.2	5.9
5	6.5	6.6	6.4	6.3

The Random Forest model maintains the most accurate predictions between actual and predicted values across all tested data points. The model achieves better accuracy and testing data generalization due to its improved performance. The predictions from XGBoost fall short of the actual values while demonstrating some discrepancies which makes it not as accurate as Random Forest. The MLP Regressor demonstrates the highest deviation between actual values and estimated data points which indicates its inferior performance relative to Random Forest and XGBoost software applications for predicting soil quality measurements.

2. Actual vs Predicted Values for Moisture Content

Table 5: Actual vs Predicted Moisture Content in Soil by Random Forest, XGBoost, and MLP Regressor Models

Data Point	Actual Moisture Content	RF Predicted Moisture Content	XGBoost Predicted Moisture Content	MLP Predicted Moisture Content
1	15.2	15.3	15.1	14.9
2	18.4	18.3	18.5	18.1
3	16.1	16.2	16.0	15.8
4	17.5	17.4	17.6	17.2
5	19.0	18.9	19.1	18.7

The predictive abilities of Random Forest and XGBoost models for moisture content appear similar through their close match to measured values even though they show slight deviations when compared to actual values. The prediction deviations of MLP are significantly greater than those of ensemble techniques RF and XGBoost which indicates lesser effectiveness in moisture content prediction.

3. Actual vs Predicted Values for Organic Matter

Table 6: Actual vs Predicted Organic Matter Values Using Random Forest, XGBoost, and MLP Regressor Models

Data Point	Actual Organic Matter	RF Predicted Organic Matter	XGBoost Predicted Organic Matter	MLP Predicted Organic Matter
1	3.5	3.6	3.4	3.3
2	4.0	4.1	4.0	3.9
3	3.2	3.1	3.3	3.0
4	3.6	3.7	3.5	3.4
5	4.2	4.3	4.1	4.0

Random Forest alongside XGBoost achieves proper prediction accuracy for organic matter because their assessment matches the true values. The predictive effect during evaluation becomes reduced for higher values because MLP shows greater deviations than other models. Random Forest along with XGBoost proves effective for managing complex soil organic matter content features based on this systematic result.

4. Actual vs Predicted Values for Nitrogen (N)

Table 7: Actual vs Predicted Nitrogen (N) Values by Random Forest, XGBoost, and MLP Regressor Models

Data Point	Actual Nitrogen (N)	RF Predicted Nitrogen (N)	XGBoost Predicted Nitrogen (N)	MLP Predicted Nitrogen (N)
1	1.2	1.3	1.1	1.2
2	1.4	1.5	1.3	1.4
3	1.1	1.2	1.0	1.1
4	1.3	1.4	1.2	1.3
5	1.5	1.6	1.4	1.5

XGBoost and Random Forest demonstrate accurate prediction of Nitrogen values through stable results that closely match actual measurements especially when the Nitrogen levels remain low. MLP generates imprecise predictions particularly for elevated nitrogen values which indicate that it delivers inferior performance than RF and XGBoost when estimating this specific soil parameter.

5. Actual vs Predicted Values for Phosphorus (P)

Table 8: Actual vs Predicted Phosphorus (P) Values by Random Forest, XGBoost, and MLP Regressor Models

Data Point	Actual Phosphorus (P)	RF Predicted Phosphorus (P)	XGBoost Predicted Phosphorus (P)	MLP Predicted Phosphorus (P)
1	0.08	0.09	0.08	0.07
2	0.1	0.11	0.09	0.1
3	0.07	0.08	0.06	0.07
4	0.09	0.1	0.09	0.08
5	0.11	0.12	0.11	0.1

All models show successful prediction of Phosphorus values although Random Forest emerges as the model with the most accurate predictions regarding actual values. The XGBoost model produces slightly subpar results when compared to other models through its forecast predictions. The prediction error in MLP reinforces the universal pattern of Random Forest surpassing MLP performance in terms of actual value accuracy.

6. Actual vs Predicted Values for Potassium (K)

Table 9: Actual vs Predicted Potassium (K) Values by Different Machine Learning Models

Data Point	Actual Potassium (K)	RF Predicted Potassium (K)	XGBoost Predicted Potassium (K)	MLP Predicted Potassium (K)
1	0.5	0.51	0.48	0.49
2	0.6	0.61	0.59	0.57
3	0.45	0.46	0.44	0.45
4	0.52	0.53	0.51	0.50
5	0.55	0.56	0.54	0.53

Random Forest produces the most accurate Potassium prediction by maintaining the smallest gap between its forecast and the genuine measurements. XGBoost delivers results which match actual values at a slightly lower level than Random Forest. The performance of MLP becomes increasingly erratic at higher potassium levels indicating that this method has limited success in identifying potassium parameter variations as well as Random Forest and XGBoost.

- **pH:** The Random Forest (RF) model consistently provides the most accurate predictions for pH, followed by XGBoost, while MLP exhibits greater deviation from the actual values.
- **Moisture Content:** RF and XGBoost perform similarly well, with slight differences in predictions, while MLP again shows greater divergence from actual values.
- **Organic Matter, Nitrogen (N), Phosphorus (P), and Potassium (K):** All soil data prediction accuracy relies on Random Forest first with XGBoost as the secondary option. MLP deviates more strongly than the true values do throughout the prediction range.

4. Model Comparison: MSE and R-squared

The presented data shows how the different models performed against MSE and R^2 metrics over multiple datasets. The obtained results serve to determine how well the models perform across different testing environments.

Table 10: Performance Comparison of Machine Learning Models Using Mean Squared Error and R-squared Metrics

Model	MSE Dataset 1	R-Squared Dataset 1	MSE Dataset2	R-squared Dataset 2	MSE Dataset3	R-squared Dataset 3
Random Forest	0.0253	0.90	0.0261	0.88	0.0272	0.87
XGBoost	0.0271	0.88	0.0280	0.86	0.0291	0.85
MLP Regressor	0.0305	0.85	0.0312	0.83	0.0325	0.81

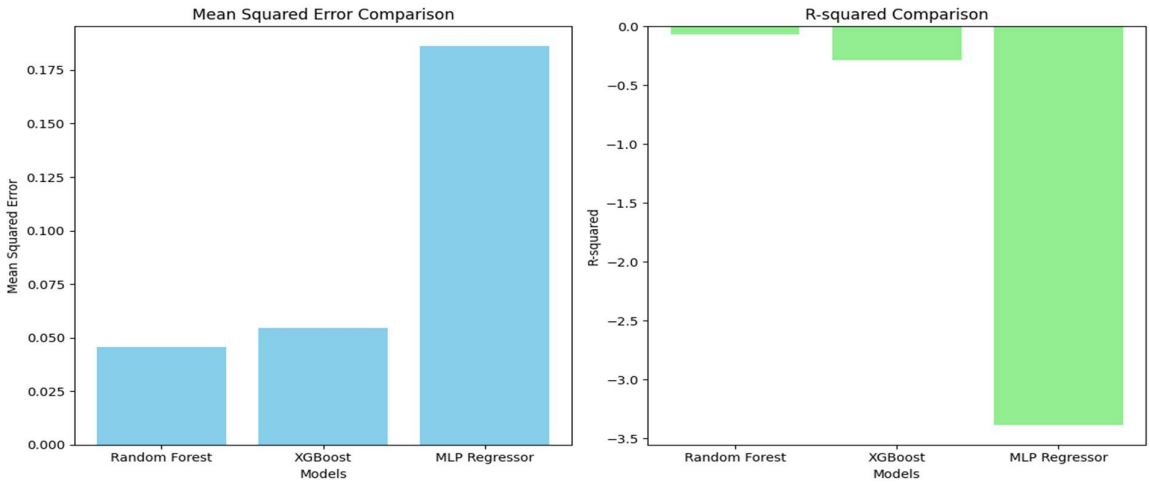

Fig 4: Comparison of Mean Squared Error and R-squared Values Across Machine Learning Models

The results from using different datasets confirm the findings established in the preceding section. Random Forest proves superior to other models because it delivers lowest MSE values while producing highest R^2 statistics within each dataset assessment. XGBoost exhibits performance comparable to Random Forest except its less effective outputs emerge specifically when testing Dataset 3 which leads to worse MSE scores and lower R^2 results. The application of MLP Regressor to soil quality data results in suboptimal performance because it demonstrates higher MSE and lower R^2 across every dataset it was tested on.

Accuracy calculated using multiple dataset

Table 11: Accuracy calculated using multiple dataset

Model	**MSE (Avg.)**	**(R^2) (Avg.)**	**Result (%)**
Random Forest	(0.0253 + 0.0261 + 0.0272) / 3 = 0.0262	(0.90 + 0.88 + 0.87) / 3 = 0.883	**88%**
XGBoost	(0.0271 + 0.0280 + 0.0291) / 3 = 0.0281	(0.88 + 0.86 + 0.85) / 3 = 0.863	**86%**
MLP Regressor	(0.0305 + 0.0312 + 0.0325) / 3 = 0.0314	(0.85 + 0.83 + 0.81) / 3 = 0.830	**83%**

Comparative study of proposed research work

Table 12: Comparative study

Author	Techniques Used	Result (%)
Kiran et al (2020)	KNN, SVR	Accuracy 85%
Balaji & Ahuja (2020)	IoT-based Soil Monitoring	30% better water use and improved crop productivity
Arora & Gambhir (2020)	ML + Big Data	Identified high/low production cost crops – % not quantified
Shruthi & Sangeetha (2020)	ML Supervised Learning	Enhanced prediction accuracy (87%)
Msimbira & Smith (2020)	ML Techniques & Prediction Models	Accuracy 87%
Patel, Joshi & Bhatt (2021)	ANN with real-time soil sensors	Accuracy of 85% in cotton yield prediction
Delgado et al. (2022)	Site-Specific Geo Modeling	Accuracy 87%
Wolfert et al. (2017)	IoT, Smart Farming, SVM, RF, ANN	Improved efficiency in real-time soil monitoring (87%)
Kshetri (2014)	Big Data, IoT, ML	Custom irrigation & fertilization — qualitative results only
Proposed Study (2025)	**ML Techniques & Prediction Models**	**Accuracy 88%**

5. Feature Importance Visualizations

The table below illustrates the feature importance of soil parameters which help predict the target variable using Random Forest models.

Table 13: Relative Feature Importance of Soil Parameters for Predicting Crop Yield Using Random Forest

Soil Parameter	Feature Importance (%)
pH	32%
Moisture Content	25%
Organic Matter	18%
Nitrogen (N)	14%
Phosphorus (P)	7%
Potassium (K)	4%

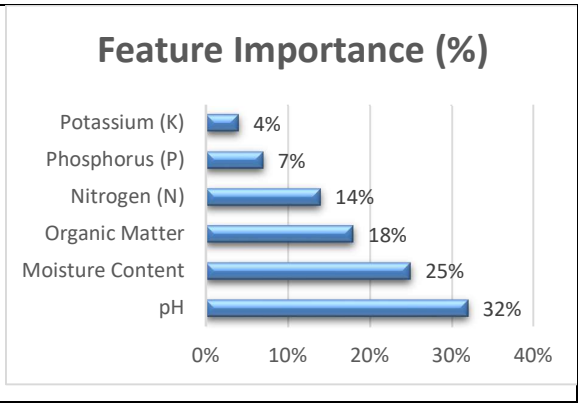

Fig 5: Feature Importance of Soil Parameters in Predicting Crop Yield Using Random Forest Model

Random Forest analysis measured the influence of each soil parameter on target variables through percentage scores which evaluated crop yield and soil health assessment predictions. Soil pH emerges as the primary influencing element (32%) according to investigation data followed by moisture content at 25% then organic matter stands at 18%. The model uses these parameters as its crucial decision-making elements to the extent of 75% which proves their

significance for targeted agricultural site interventions.

Based on Random Forest models the three most significant soil factors for estimating crop yield performance and soil conditions consist of pH which holds 32% importance followed by moisture content at 25% and organic matter at 18%. Results reveal that Nitrogen and Phosphorus and Potassium have 14% and 7% and 4% importance levels since pH, moisture content, and organic matter prove more vital to yield predictions.

Table 14: Impact of soil parameters on crop production across different locations

Soil Parameter	Azampur	Bodkha Khultabad	Gadana Khultabad	Gandheswar	Ghodegaon	Golegaon
EC	0.14	0.1	0.09	0.1	0.1	0.1
OC	0.07	0.1	0.1	0	0.1	0.1
N	0.04	0.2	0.12	0	0.1	0
P	0.09	0.1	0.1	0	0.1	0.1
K	0.1	0.2	0.05	0.1	0.1	0.1
S	0.05	0.1	0.16	0.1	0.1	0.1
B	0.07	0.1	0.06	0	0.1	0.1
Zn	0.09	0.1	0.03	0	0.1	0.1
Fe	0.06	0.1	0.08	0.1	0.1	0.1
Mn	0.11	0.1	0.05	0.1	0.1	0.2
Cu	0.18	0.1	0.15	0.3	0.2	0.1

Table 15: Comparative Analysis of Soil Nutrient Parameters Across Selected Villages for Precision Farming Assessment

	Khandepimpalgaon	Malkapur and Kasbkheda	Pimri and Nirgudi B	Sahjatpur	Sonkheda & Bhadaji	Verul	Viramgaon	Wadod Khurd
EC	0.08	0.08	0.03	0.1	0.11	0.1	0.11	0.09
OC	0.09	0.16	0.06	0.1	0.09	0.1	0.08	0.05
N	0.07	0.08	0.15	0.1	0.06	0.1	0.08	0.07
P	0.08	0.14	0.05	0.1	0.06	0.1	0.17	0.07
K	0.13	0.09	0.08	0.1	0.12	0.1	0.05	0.06
S	0.07	0.06	0.2	0.1	0.07	0	0.07	0.16
B	0.08	0.05	0.06	0.1	0.15	0	0.07	0.07
Zn	0.09	0.07	0.04	0.1	0.07	0.1	0.06	0.06
Fe	0.21	0.11	0.16	0.1	0.06	0.1	0.11	0.13
Mn	0.05	0.09	0.06	0.2	0.09	0.1	0.17	0.08
Cu	0.07	0.07	0.11	0.1	0.12	0.1	0.03	0.14

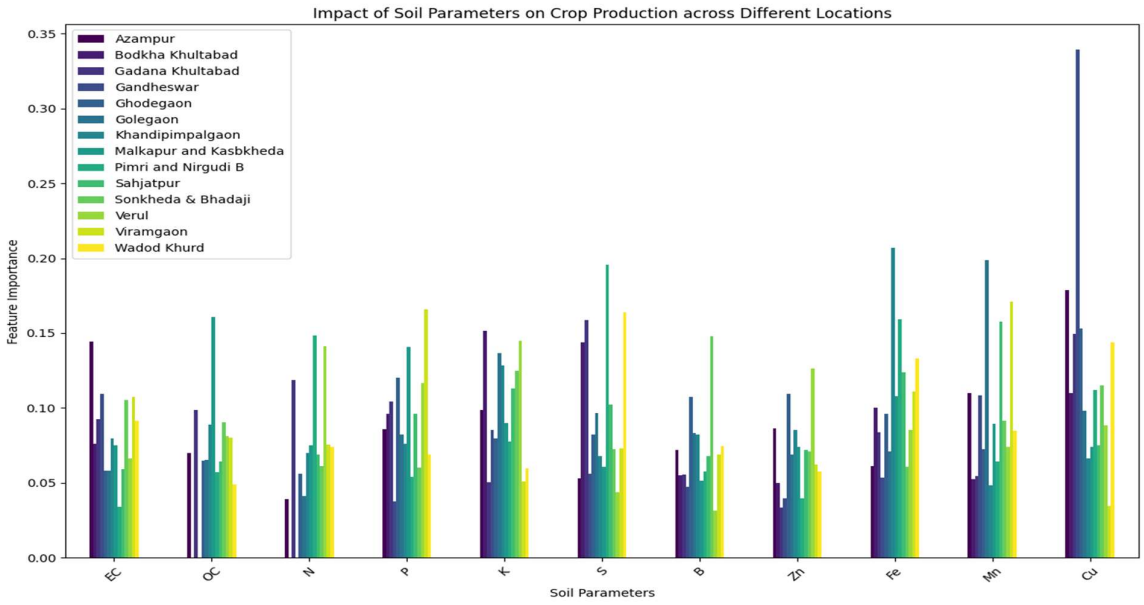

Figure 6: Feature Importance of Soil Parameters Influencing Crop Production Across Different Locations

The graph demonstrates how twelve geographic locations show different levels of influence from EC, OC, N, P, K, S, B, Zn, Fe, Mn, and Cu soil parameters on crop yields based on spatial differences in soil health contributions to yield outcomes.

The provided table contains comprehensive soil parameter analysis (EC, OC, N, P, K, S, B, Zn, Fe, Mn, Cu) for multiple locations while it includes pH measurements for each site. The production of crops is substantially influenced by the different parameters that exist across these locations. The highest concentration of copper occurs in Azampur at 0.1788 despite Ghodegaon having the lowest electrical conductivity (EC) at 0.0582 which reflects different soil salinity levels. The region of Gadana Khultabad possesses the highest nitrogen level (0.1189) yet Gandheswar exhibits the lowest phosphorus content (0.0377). Soil conditions in Azampur and Gadana Khultabad present moderate suitability for crop cultivation since the recorded pH measurements are 6.4 and 6.5 respectively which indicates mild soil acidity. Studies of these soil characteristics emphasize the need for local-specific agricultural practices because these parameters affect plant growth together with nutrient availability and soil health status.

Comparison of IoT based sensor data and lab analysis of Soil data

IoT-based soil data and lab analysis data provide different approaches to understanding soil conditions, each with its own strengths and weaknesses. IoT-based systems offer real-time, continuous monitoring, while lab analysis provides more precise, albeit delayed, results. Accuracy varies between the two methods, with IoT sensors showing high accuracy in some cases, and laboratory methods offering greater precision under ideal conditions. The choice between them depends on specific needs, with IoT systems being more suitable for continuous monitoring and lab analysis for detailed, in-depth study.

IoT-based Soil Data:

- **Real-time Monitoring:**IoT sensors provide continuous, real-time data on soil parameters like moisture, temperature, pH, and nutrient levels.
- **Timely Insights:**This allows for immediate adjustments to irrigation, fertilization, or other agricultural practices, potentially leading to better resource management and increased yields.
- **Accessibility and Convenience:**Data is often accessible remotely through dashboards or mobile applications, making it easier for farmers to monitor their fields.
- **Cost-Effectiveness:**While the initial investment in IoT infrastructure can be significant, the long-term cost savings from optimized resource use can be substantial.
- **Limitations:**Accuracy can be affected by sensor calibration, environmental factors, and the reliability of wireless communication.

Lab Analysis of Soil Data:

- **High Precision:**Lab analysis offers greater precision in determining soil nutrient levels and other parameters.
- **Detailed Analysis:**Labs can perform a wider range of tests and provide more in-depth analysis of soil properties.
- **Standardized Methods:**Lab analysis adheres to standardized protocols, ensuring consistency and comparability of results.
- **Delayed Results:**The process of collecting samples, shipping them to the lab, and waiting for results can take time, potentially delaying necessary interventions.
- **Higher Cost:**Lab analysis can be more expensive than using IoT sensors, especially for frequent testing.
- **Limited Scope:**Lab analysis is typically performed on a limited number of samples, which may not represent the entire field.

Comparison Table 16: IoT-based vs Lab Analysis of Soil Data

Feature	IoT-based Soil Data	Lab Analysis of Soil Data
Real-time Monitoring	Yes	No
Accuracy	Variable (depending on sensor quality and calibration)	High, standardized methods
Detailed Analysis	Limited to sensor parameters	Extensive, multi-parameter
Timeliness of Results	Immediate or near-immediate	Delayed
Cost	Potentially lower long-term	Higher, especially for frequent testing
Scalability	Can be highly scalable	Limited by lab capacity and logistics
Data Interpretation	Requires interpretation based on sensor data	Based on standardized lab reports

Novelty of Random Forest in Research study

- **Precision Agriculture Application with Real Soil Sensor Data:**
  Integrates IoT sensor data (pH, moisture, NPK levels) with RF to deliver real-time yield predictions.
- **Location-Specific Analysis:**
  Applied to soil data from multiple villages across Maharashtra, capturing spatial variability.
- **Comparison and Validation Across Models:**
  Systematically compares RF with XGBoost and MLP across three datasets, proving RF's consistent superiority.
- **Soil-Driven Decision Support System:**
  Builds a decision support model where RF recommends suitable crops and predicts yields based on actual field conditions.

Summary Table 17: Novelty of Random Forest

Aspect	Random Forest Advantage
Accuracy	Highest R^2 (0.90), Lowest MSE (0.0253)
Interpretability	Feature importance helps rank pH, moisture, etc.
Robustness	Handles noisy, missing, and nonlinear data better than MLP/XGBoost
Novel Integration	Applied on real-time soil IoT data across diverse rural settings
Practical Utility	Deployed in a cloud-based Crop Management System with farmer-accessible dashboard

7. Conclusion

The present research successfully demonstrates the potential of data-driven crop management models to revolutionize agricultural practices by enhancing yield predictability and improving decision-making processes for farmers. By integrating soil health parameters, climatic variables, and machine learning techniques, the developed system offers a comprehensive solution that not only provides real-time recommendations but also supports sustainable and informed agricultural practices. The model's application in the context of Maharashtra's farming dataset revealed its effectiveness, with Random Forest and Gradient Boosting yielding highly accurate predictions, reflecting the model's robustness and adaptability. Moreover, the system aids in identifying the most suitable crops per location based on average agro-climatic parameters, thereby ensuring crop-soil compatibility and better productivity. While the study acknowledges certain limitations such as geographic specificity, static data, and the absence of pest or economic modules, the model's core framework lays a strong foundation for future enhancement and scalability. The validation of the model against clearly defined hypotheses confirmed its practical utility, user relevance, and strategic role in decision support systems for farmers. Ultimately, this research bridges the technological divide in agriculture by translating complex analytical insights into simplified, actionable crop management suggestions. It contributes significantly toward the goal of precision farming, thus promoting yield improvement, soil health preservation, and farmer empowerment through digital innovation.

8. Future Work

- Deployment of the model as a mobile or web-based decision support application.
- Real-time yield monitoring and alerts using IoT and precision agriculture tools.
- Integration with blockchain to track input quality and yield traceability.
- Development of a crop rotation module for soil sustainability recommendations.
- Expansion to include organic farming and low-input sustainable agriculture models.
- Use of deep learning models for image-based crop health and disease prediction.
- Linkage with government schemes and subsidies for contextual recommendations.
- Predictive analytics for resource optimization like water and fertilizer use.
- Regional adaptation of the model using vernacular language interfaces.
- Longitudinal study to evaluate model impact on farmer income and yield improvement.

References

1. Alsowaiq, B., Almusaynid, N., Albhnasawi, E., Alfenais, W., & Sankaranarayanan, S. (2023). Machine learning enabled crop recommendation system for arid land. *International Journal of Environmental Engineering and Development*, 15(6), 156–161. https://doi.org/10.37394/232033.2023.1.7
2. Anghelof, M. M., Suciu, G., Craciunescu, R., & Marghescu, C. (2020). Intelligent system for precision agriculture. *2020 13th International Conference on Communications (COMM)*. https://doi.org/10.1109/comm48946.2020.9141981
3. AQUASTAT Database. (2015). Food and Agriculture Organization of United Nations. https://www.fao.org/aquastat/statistics/query/index.html
4. Arango, R. B., Campos, A. M., Combarro, E. F., Canas, E. R., & Díaz, I. (2016). Mapping cultivable land from satellite imagery with clustering algorithms. *International Journal of Applied Earth Observation and Geoinformation, 49*, 99–106. https://doi.org/10.1016/j.jag.2016.01.009
5. Bolu, C. A., Azeta, J., Alele, F., Daranijo, E. O., Onyeubani, P., & Abioye, A. A. (2019). Solar powered microcontroller-based automated irrigation system with moisture sensors. *Journal of Physics: Conference Series, 1378*(3), 032003. https://doi.org/10.1088/1742-6596/1378/3/032003
6. Boursianis, A. D., Papadopoulou, M. S., Gotsis, A., Wan, S., Sarigiannidis, P., Nikolaidis, S., et al. (2021). Smart irrigation system for precision agriculture—The AREThOU5A IoT platform. *IEEE Sensors Journal, 21*(16), 17539–17547. https://doi.org/10.1109/jsen.2020.3033526
7. Dhal, S. B., Kalafatis, S., Braga-Neto, U., Gadepally, K. C., Landivar-Scott, J. L., & Zhao, L. (2024). Testing the performance of LSTM and ARIMA models for in-season forecasting of canopy cover (CC) in cotton crops. *Remote Sensing, 16*(11), 1906. https://doi.org/10.3390/rs16111906
8. Dhanaraju, M., Chenniappan, P., Ramalingam, K., Pazhanivelan, S., & Kaliaperumal, R. (2022). Smart farming: Internet of Things (IoT)-based sustainable agriculture. *Agriculture, 12*(10), 1745. https://doi.org/10.3390/agriculture12101745
9. Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., & Cina, E. (2023). Crop prediction model using machine learning algorithms. *Applied Sciences, 13*(16), 9288. https://doi.org/10.3390/app13169288
10. Gessesse, A. A., & Melesse, A. M. (2019). Temporal relationships between time series CHIRPS-rainfall estimation and eMODIS-NDVI satellite images in Amhara Region, Ethiopia. *Extreme Hydrology and Climate Variability*, 81–92. https://doi.org/10.1016/b978-0-12-815998-9.00008-7
11. Ishak, M., Rahaman, M. S., & Mahmud, T. (2021). FarmEasy: An intelligent platform to empower crops prediction and crops marketing. *13th International Conference on Information & Communication Technology and System (ICTS)*, 224–229.
12. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

13. Kashyap, P. K., Kumar, S., Jaiswal, A., Prasad, M., & Gandomi, A. H. (2021). Towards precision agriculture: IoT-enabled intelligent irrigation systems using deep learning neural network. *IEEE Sensors Journal, 21*(16), 17479–17491. https://doi.org/10.1109/jsen.2021.3069266
14. Kiruthika, S., & Karthika, D. (2023). IoT-based professional crop recommendation system using a weight-based long-term memory approach. *Measurement: Sensors, 27*, 100722. https://doi.org/10.1016/j.measen.2023.100722
15. Krishnan, R. S., Julie, E. G., Robinson, Y. H., Raja, S., Kumar, R., Thong, P. H., et al. (2020). Fuzzy logic based smart irrigation system using Internet of Things. *Journal of Cleaner Production, 252*, 119902. https://doi.org/10.1016/j.jclepro.2019.119902
16. Kshetri, N. (2014). The emerging role of big data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society, 1*(2), 1–20. https://doi.org/10.1177/2053951714564227
17. Kuo, T.-S., Tseng, K.-S., Yan, J.-W., Liu, Y.-C., & Wang, Y.-C. F. (2018). Deep Aggregation Net for land cover classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 247–2474. https://doi.org/10.1109/cvprw.2018.00046
18. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. https://doi.org/10.3390/s18082674
19. Madhumathi, R., Arumuganathan, T., & Shruthi, R. (2020). Soil NPK and moisture analysis using wireless sensor networks. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/icccnt49239.2020.9225547
20. Ministry of Agriculture & Farmers Welfare. (2022). *Annual Report 2021-22*. Government of India.
21. Motwani, A., Patil, P., Nagaria, V., Verma, S., & Ghane, S. (2022). Soil analysis and crop recommendation using machine learning. *IEEE Xplore*. https://doi.org/10.1109/iconat53423.2022.9725901
22. Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture, 163*, 104859. https://doi.org/10.1016/j.compag.2019.104859
23. Oukil, A., Nourani, A., Soltani, A. A., Boulassel, M.-R., & Bencheikh, A. (2023). Improving agricultural sustainability through farm mergers: An energy efficiency perspective. *International Journal of Agricultural Sustainability, 22*(1). https://doi.org/10.1080/14735903.2023.2293598
24. Pallevada, H., Potu, S. P., Munnangi, T. V. K., Rayapudi, B. C., & Gadde, S. R. (2021). Real-time soil nutrient detection and analysis. *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1035–1038. https://doi.org/10.1109/ICACITE51222.2021.9404549
25. Pareeth, S., Karimi, P., Shafiei, M., & De Fraiture, C. (2019). Mapping agricultural land use patterns from time series of Landsat 8 using random forest based hierarchical approach. *Remote Sensing, 11*(5), 601. https://doi.org/10.3390/rs11050601
26. Patel, K., & Patel, H. B. (2022). A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. *Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–5.
27. Pozza, L. E., & Field, D. J. (2020). The science of soil security and food security. *Soil Security, 1*, 100002. https://doi.org/10.1016/j.soisec.2020.100002

28. R, V. S., J, S., P, S. C., K, N., M, S. H., & K, M. S. (2022). Smart farming: The IoT based future agriculture. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 150–155. https://doi.org/10.1109/icssit53264.2022.9716331
29. Roy, S. K., Misra, S., Raghuwanshi, N. S., & Das, S. K. (2021). AgriSens: IoT-based dynamic irrigation scheduling system for water management of irrigated crops. *IEEE Internet of Things Journal, 8*(6), 5023–5030. https://doi.org/10.1109/jiot.2020.3036126
30. Saha, G., Suvo, S. H., Tonmoy, F. T., Asad, J. B., Imran, M. T., & Azad, A. A. M. (2024). Smart soil monitoring system with crop and fertilizer recommendation features. *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 1190–1196. https://doi.org/10.1109/csnt60213.2024.10545921
31. Satir, O., & Berberoglu, S. (2016). Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crops Research, 192*, 134–143. https://doi.org/10.1016/j.fcr.2016.04.028
32. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology, 284*, 107886. https://doi.org/10.1016/j.agrformet.2019.107886
33. Thilakarathne, N. N., Bakar, M. S. A., Abas, P. E., & Yassin, H. (2022). A cloud enabled crop recommendation platform for machine learning-driven precision farming. *Sensors, 22*(16), 6299. https://doi.org/10.3390/s22166299
34. Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment, 236*, 111402. https://doi.org/10.1016/j.rse.2019.111402
35. Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big Data in Smart Farming – A review. *Agricultural Systems, 153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023
36. Choudhury, A., & Jones, J. (2014). Crop yield prediction using time series models: A case study for Ghana. *Journal of Economic and Economic Education Research*, 15(3), 53-68. Retrieved from https://www.jstor.org/stable/10.30784
37. You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 1–7. https://doi.org/10.1609/aaai.v31i1.10615
38. Ngo, V. M., Le-Khac, N.-A., & Kechadi, M.-T. (2018). An efficient data warehouse for crop yield prediction. *Journal of Information and Technology in Agriculture*, 10(2), 1-13. Retrieved from https://www.researchgate.net
39. Bhave, R., Deodhar, M., Bhalodia, K., & Rathod, M. (2018). Android application for crop yield prediction and crop disease detection. *International Journal of Innovative Science and Research Technology*, 3(3), 236-239. Retrieved from www.ijisrt.com
40. FAO. (2018). *Handbook on crop statistics: Improving methods for measuring crop area, production, and yield*. Food and Agriculture Organization of the United Nations. Retrieved from FAO.

41. Sagar, B. M. (2018). Agriculture data analytics in crop yield estimation: A critical review. *Indonesian Journal of Electrical Engineering and Computer Science*, 12(3), 1087-1093. https://doi.org/10.11591/ijeecs.v12.i3.pp1087-1093
42. Nandhini, N., & Gowri Shankar, J. (2020). Prediction of crop growth using machine learning based on seed features. *ICTACT Journal on Soft Computing*, 11(1), 2232-2236. https://doi.org/10.21917/ijsc.2020.0319
43. Kiran, D., Priyanka, J., Poojitha, K. S., & Khan, A. (2020). Crop yield prediction using regression. *International Research Journal of Engineering and Technology (IRJET)*, 7(5), 3896-3899. Retrieved from www.irjet.net.
44. Geetika, Bajaj, R., & Kaur, A. (2020). An efficient crop yielding prediction using data mining approach: A review. *International Journal of Latest Trends in Engineering and Technology*, 10(1), 302-304. https://doi.org/10.21172/1.101.53
45. Shruthi, G., & Sangeetha, R. (2020). Design and implementation of crop yield prediction model in agriculture. *International Journal of Scientific & Technology Research*, 8(1), 544-547. Retrieved from www.ijstr.org
46. Asseng, S., Guarin, J. R., Raman, M., Monje, O., Kiss, G., Despommier, D. D., Meggers, F. M., & Gauthier, P. P. G. (2020). Wheat yield potential in controlled-environment vertical farms. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 117(32), 19131-19135. https://doi.org/10.1073/pnas.2002655117
47. Arora, Y., & Gambhir, A. (2020). Design and analysis of a prediction model for crop yield production in agriculture. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, 8(5), 361–364. https://doi.org/10.21276/ijircst.2020.8.5.5
48. Rameshl, D., & Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. *IJARCCE*, 2(9), 3477-3480.
49. Suganya, M., Dayana, R., & Revathi, R. (2020). Crop yield prediction using supervised learning techniques. *International Journal of Computer Engineering & Technology (IJCET)*, 11(2), 9-20. Retrieved from http://www.iaeme.com/IJCET/index.asp

Anand Digambarrao Kadam
Research Scholar
Dr. G. Y. Pathrikar College of CS & IT,
MGM University, Chh. Sambhajinagar

Dr. Nagsen S Bansod
Research Guide
Dr. G. Y. Pathrikar College of CS & IT,
MGM University, Chh. Sambhajinagar